ISSN 0971-0388

Aligarh Journal of Statistics
Vol. 36 (2016), 93-116

Multivariate Extension Of Generalized Linear Model For Polytomous Data : A Bayes Study

Richa Srivastava[1] , S.K. Upadhyay[2] and V. K. Shukla[3]

[Received on March, 2015. Revised on June, 2016]

ABSTRACT

The paper provides a multivariate extension of generalized linear model for polytomous data and considers the logistic regression model as a special case. Complete Bayes analysis of polytomous data assuming logistic regression model is provided using appropriate non-informative priors for the parameters. Since the resulting posteriors analysis is quite complex, appropriate Markov chain Monte Carlo algorithm has been developed for the same. Results are illustrated on the basis of a real data example related to biliary acid constituents of the patients having gallbladder diseases.

1. INTRODUCTION

An epidemiological research comprises two types of studies depending on whether the events have already happened (retrospectively) or whether the events may happen in the future (prospectively). The most common studies are the retrospective studies which are also called case-control studies. Case-control study is an analytical study in which a group of patients having a particular disease (cases) is compared with a group of persons who do not have that disease (controls) but exposed to the risk of the same. Study is done with respect to the exposure of one or more than one risk factors. The

[1] *Address for Correspondence:* Richa Srivastava, Jaipuria Institute of Management, Lucknow, India. E-mail: *rsrivastava.bhu@gmail.com,* S.K. Upadhyay[2], Department of Statistics, BHU, Varanasi and V. K. Shukla[3], Department of General Surgery, Institute of Medical Sciences, Banaras Hindu University, Varanasi, India.

central theme of a case-control study is to compare the diseased group with non-diseased group with respect to the exposure of risk factor. This type of study provides an association of the risk factor with the (often) higher incidence of disease in the population exposed to that particular risk factor. The measurement of association between the exposure and occurrence of the disease is done through odds ratio which is the ratio of odds of exposure in diseased group to that of non-diseased group.

Odds ratio is an important measure of relative risk that ranges from zero to infinity in value. A value close to unity indicates no relationship between the occurrence of the disease and the exposure risk factor. On the other hand, a value less than unity or a value greater than unity indicates the protective or causative effect of the exposure factor, respectively. Truly speaking, odds ratio is the answer to the question, how frequently the exposure to a risk factor is present in each group of cases and controls to determine the relationship between the risk factor and the disease. This measure is directly connected to the logistic regression relationship for dichotomous or binary responses that models the natural logarithm of odds ratio as the linear function of the predictor variables and also provides the possibility to generalize the odds ratio beyond dichotomous responses.

It is worth mentioning that several types of responses are encountered by the biomedical researches and, therefore, it is very common to categories the patients on the basis of their responses to any treatment, severity of disease or some test results, etc. Suppose, for instance, the response of a patient is denoted by a dichotomous or binary type random variable Y and, as such, Y takes value either unity or zero. This actually reflects two categorizations of the patients such as alive or dead, cured or uncured, injured or not injured, HIV positive or HIV negative, etc. Besides, there may be some intermediate or adjacent categories (or responses) apart from the two main categories. Say, for instance, a patient can be categorized according to the severity of the disease as ‘mild’, ‘modest’ or ‘severe’. Similarly, a doctor can categorize a patient as ‘good’, ‘fair’, ‘serious’ or ‘critical’ based on the stages of (say) cancer growth. All such instances or categories can be treated as polytomous where the variable Y can be assigned several values, say 0, 1, 2, ... based on the types of responses or categorizations.

Generally, polytomous responses are classified into three different categories in accordance with three different scales. These are often termed as nominal, ordinal and interval scales. The nominal scales are those where categories can be interchanged and there is no specific ordering among the different categories. An important example can be the categorization according to (say) eye’s colour as ‘black’, ‘brown’, ‘blue’. Nominal scales are generally treated as the lowest scales of polytomous responses. In ordinal scale, categories can be placed in a specific manner much like the ordinal numbers (either in ascending or in descending order) and they are not exchangeable. Examples of such polytomous responses include degree of injuries and tumor sizes, stages of cancer, etc. It is to be noted that in such categorizations, it does not make sense to talk of the notion of distance and spacing among the categories. Say, for example, it is wrong to say that a particular category is three times greater (or less) than the other one. Third type of polytomous responses include interval scales where the categories can be ordered and numerical labels can also be attached. Differences between the numerical labels (or scores) on different categories can, therefore, be interpreted as a measure of separation of the categories. Examples of responses with interval scales include age distribution, blood pressure levels, pulse counts, etc. where scores are generally defined at the midpoint of the intervals.

The responses resulting to polytomous data may incorporate two kinds of correlations among the variates. This may be because of the fact that responses to various polytomous items may be dependent as a consequence of clustered or mingled structure of data and also because of the fact that the response to a single polytomous item can be seen as a multivariate dependent response. These two correlations are often referred to as within cluster and between cluster correlations and play a crucial role while defining a model for polytomous data. There may be situations, however, when one or both the correlations may not appear directly or may be weak enough to foresee. Literature provides enough references on various kinds of modelling related to different response variables as well as their analyses. A few among these can be cited as Agresti (2002), McCullagh and Nelder (1989), Liang (1999), etc.

For a dichotomous response with $Y = 1$ or 0, we obviously get among others the logistic regression model, which is a special case of generalized linear model (GLM) with “logit” link and binomial distribution for the random component Y, the probability of success being the expectation of Y . GLM has its natural extension for polytomous data as well and the resulting special case can be referred to as the polytomos logistic regression where the link function as usual is “logit” but the distribution of random component is now multivariate Bernoulli, which is same as a multinomial distribution with total count unity. Generally speaking, a polytomous variable can be structured as the multivariate variable and, among other things, the components of this multivariate variable may be taken as binary type having some correlation among them.

Both the multivariate extension of GLM and its special case, the polytomous logistic regression model, has been considered extensively in the classical statistical literature. Rasch (1961) is perhaps an early reference that provides the applicability of one such model in an educational measurement context although GLM was first introduced by Nelder and Wedderburn (1972). Other significant references include Andrich (1978), Agresti (2002), etc., where both GLM and its different special cases are successfully dealt. It is to be noted that GLM is one of the most important contributions that unifies several models such as multiple linear regression, logistic regression, log-linear regression and Poisson regression, among others. Since these latter models are regularly employed by various researchers to study how the values of outcome variable vary over the different configuration of predictor variables, GLM has given an easy response where all these models can be housed together and successfully analyzed. This is perhaps the reason that GLM has maximum applicability in almost every area including the biomedical researches. It is worth mentioning that the objective of any regression analysis including GLM in biomedical applications is to establish a kind of linking between the particular test results and disease as the study of such linking enables the researchers to understand and identify the causal factor of the level of disease so that the appropriate treatment can be accordingly suggested.

Bayesian inferences to multivariate GLM or its special case ‘polytomous logistic regression model’ are comparatively meager though the previous two decades provided significant developments in several applied areas. A few important references include Draper and Smith (1998), Congdon (2004), Gelman and Hill (2007), etc. where a few among these provide realistic examples from a variety of areas. The reason for such rapid developments can, of course, be attributed to enormous progresses in Bayesian modelling and computational techniques.

The paper is organized as follows. In Section 2, we briefly discuss the polytomous logistic regression model and provide its complete Bayesian formulation by choosing non-informative priors for its parameters. We also provide a brief discussion on the associated computational issues using Gibbs sampler algorithm. Section 3 discusses a few important Bayesian tools, namely the Bayesian information criterion (BIC), deviance information criterion (DIC) and posterior Bayes factor. This is being done for the completeness of the paper. Section 4 provides a brief description of a real data on biliary acid constituents of gallbladder patients that has been used for the purpose of numerical illustration. The corresponding numerical illustration is given in Section 5. Section 6 focuses on the justification of modeling assumption by an informal procedure and also provides the comparison of full and reduced models by using the tools discussed in Section 3. Finally, a brief conclusion is given in the last section.

2. MODELLING FORMULATION

Without going into the details of GLM or its multivariate analogue, let us focus our attention on the special case of polytomous logistic regression model. The interested readers may refer to Liang (1999), Tuerlinckx and Wang (2004), etc. for a thorough discussion on GLM, its multivariate analogue and several other special cases of GLM.

2.1 Polytomous logistic regression model

To begin with, let us consider k possible response categories (j = 0, 1, ⋯, k-1) and suppose the response of an individual falls in one of these categories. We write $Y = j$ where Y is used to denote the polytomous response variable and j is used to denote the response of an individual, j = 0, 1, ⋯, k-1. The assumption of response to lie in one of the categories simplifies the modelling formulation although one can consider a more generalized situation and allow the response of an individual to fall in more than one category. Clearly, k=2 refers the binary case where Y is allowed to take only two values 0 and 1. Moreover, the difference between binary and polytomous data is that the latter is multivariate or a vector valued random variable. In order to provide basis for defining an appropriate model, we further associate a random vector $\boldsymbol{C}$ consisting of $k-1$ components (the length is one less than the number of categories) with each response Y=j and assign values zeros and ones to the components of $\boldsymbol{C}$. If c_j, $j = 1,2,...,k-1$ denotes the j^{th} component of this random vector $\boldsymbol{C}$, it may be defined as

$$c_j = \begin{cases} 1, if \quad Y = j,\ j = 1, 2, \dots, k-1 \\ 0, otherwise \end{cases} \tag{2.1}$$

More precisely, responses to different categories can be converted into a random vector as shown in Table 1.

Table 1: A simple illustration of polytomous data

Response variable *Y*	Random Vector ***C***				
	c_1	c_2	c_3	...	c_{k-1}
0	0	0	0	...	0
1	1	0	0	...	0
2	0	1	0	...	0
.	.	.	.	...	.
k-1	0	0	0	...	1

Obviously, Table 1 provides the indicator version of polytomous responses of the variable *Y* which may be used to define the probability mass function given as

$$P(Y=j) = p_1^{c_1} p_2^{c_2} \dots p_{k-1}^{c_{k-1}} (1 - p_1 - \dots - p_{k-1})^{1 - c_1 - \dots - c_{k-1}};\ j = 0, 1, \dots, k-1 \tag{2.2}$$

where Y = 0 denotes the base line category response and the probability of responding in this category is $P(Y = 0) = 1 - p_1 - p_2 ... - p_k{-}_1$. Equation (2.2) is the probability mass function of multivariate Bernoulli distribution with p_j as the probability of responding in the category j, $j = 1, ..., k - 1$, $p_0 = P(Y = 0)$ and $\sum_{j=0}^{k-1} p_j = 1$. The mean of the distribution is the vector of marginal probabilities p= $(p_1, p_2, ..., p_{k-1})^T$ and the variance of each univariate component is $p_j (1 - p_j)$, $j = 1,2,...,k - 1$. Also the covariance between two components is given by $-p_i\, p_j$, $i \neq j$.

2.2 Link function

In order to complete the modelling formulation, let us next consider defining the link function so that the response variable can be linked to the corresponding predictor variables. Here it is logical to assume that the vector-valued link function $f_{\ link}$ transforms the vector of means of the multivariate Bernoulli distribution (2.2) and it may written as

$$f_{link}(p) = (f_{link1}(p_1)\, f_{link2}(p_2) ... f_{link(k-1)}(p_{k-1}))^T = (\eta_1 \eta_2 ... \eta_{k-1})^T \qquad (2.3)$$

where $f_{\ link\ j}(\ p_{\ j})$ corresponds to $\eta_{\ j}$ and $\eta_{\ j}$ is the j^{th} linear predictor. We are using here the base line category for defining the logit link function, which is given by

$$\log \frac{p_j}{p_0} = X^T \beta_j \text{ j=1, 2, …., k-1} \qquad (2.4)$$

where X^T denotes the vector of predictor variables, $\eta_{\ j} = X^T \beta_j$ is the j^{th} linear predictor and β_j is the column matrix of intercept and regression coefficients for j^{th} category. That is

$$X = (1 \quad X_1 \quad X_2 \quad \ldots \quad X_m)^T \qquad (2.5)$$

$$\beta_j = (\beta_{0j} \quad \beta_{1j} \quad \ldots \quad \beta_{mj})^T \qquad (2.6)$$

Provided there are m predictor variables X_1, X_2, ..., X_m and j = 1, 2,$\cdots$, $k - 1$.

Solving these k-1 equations simultaneously, the probability for responding in j^{th} category can be obtained as

$$p_j = \frac{\exp(X^T \beta_j)}{1 + \sum_{j=1}^{k-1} \exp(X^T \beta_j)} \quad ; j = 1, 2, \ldots, k-1 \qquad (2.7)$$

whereas the corresponding probability for responding in base line category p_0 can be given as

$$p_0 = \frac{1}{1 + \sum_{j=1}^{k-1} \exp(X^T \beta_j)} \qquad (2.8)$$

obviously, the model formulation given above reduces to that for dichotomous data if only two categories are entertained and this may be done by collapsing the k (k ≥ 3) categories into two. In this case the response variable *Y* takes value 1 or 0 and the corresponding probability mass function is given by univariate Bernoulli distribution defined as

$$P(Y=j)=p^{j}(1-p)^{1-j} \quad ;j=0,1 \tag{2.9}$$

where *p* is the probability that *Y* takes value unity. The vector valued link function converts into a single link which transforms expected value of the response variable as a function of linear predictor X^T β. That is

$$f_{link}(p)=X^{T}\beta \Rightarrow p=f_{link}^{-1}(X^{T}\beta)$$

with logit link given as

$$\log\frac{p}{1-p}=X^{T}\beta \Rightarrow p=\frac{\exp(X^{T}\beta)}{1+\exp(X^{T}\beta)} \tag{2.10}$$

where X = (1 X_1 $X_2 \ldots X_m)^T$ and β = $(\beta_0 \;\; \beta_1 \;\; \beta_2 \;\; \ldots \; \beta_m)^T$. It may be noted that the use of subscript *j* is not needed in case of dichotomous response variable.

2.3 Bayesian modelling formulation and Gibbs sampler implementation

Let us consider a situation where n items or individuals are giving responses to be categorized in one out of k categories as detailed in Subsection 2.1. We further suppose that the vector of predictor variables is determined for each of n individuals separately. The likelihood corresponding to polytomous logistic regression model ((2.2)-(2.6)) can be written as

$$L(\underline{Y}/\beta_1,\beta_2,\ldots,\beta_{k-1},X)=\prod_{i=1}^{n}\frac{(\exp(X_i^T\beta_1))^{c_{i1}}\ldots(\exp(X_i^T\beta_{k-1}))^{c_{ik-1}}}{1+\sum_{j=1}^{k-1}\exp(X_i^T\beta_j)} \tag{2.11}$$

where $X^T{}_i$ = (1 X_{1i} X_{2i} ... X_{mi}) and X_{1i}, X_{2i} ,⋯ ,X_{mi} are the values of m predictor variables corresponding to i-th item or individual. Also each β_j, *j* = 1,2,⋯ , $k-1$, is a column matrix defined as in (2.6).

The above formulation clearly shows that the number of parameters $(k-1)$ $(m+1)$ is usually large and increases with the increasing number of response categories *k* and/or the increasing number of predictor variables *m*. With increasing *k* and *m*, the parameters are difficult to estimate by usual classical methods. To proceed in a Bayesian framework, we begin by defining

independent uniform priors for each of the components of the parameter vector β$_j$. For each component, the same may be defined separately as $G\,(\beta_{lj}) = constant\,;\; U_{lj} \le \beta_{lj} \le V_{lj},$ (2.12)
where $j = 1,2,\cdots,k - 1.$ and $l = 0,1,\cdots,m$. Since we do not have any information about the intercepts and regression coefficients, in general, we propose to consider large difference between U_{lj} and V_{lj} so that priors remain vague and the inferences may be mostly data driven.

Combining (2.11) and (2.12) via Bayes theorem yields the joint posterior distribution that can be specified up to proportionality as

$$P(\beta_1, \beta_2, \ldots, \beta_{k-1} \,|\, \underline{Y}, X) \propto \prod_{i=1}^{n} \frac{(\exp(X_i^T \beta_1))^{c_{i1}} \ldots (\exp(X_i^T \beta_{k-1}))^{c_{ik-1}}}{1 + \sum_{j=1}^{k-1} \exp(X_i^T \beta_j)} \tag{2.13}$$

Although the posterior distribution is complicated to solve analytically, the numerical solution is always an option. Among various possibilities, we can implement Gibbs sampler algorithm, a Markovian updating scheme, to simulate from the posterior (2.13). The algorithm provides a straightforward solution based on simulating from various full conditionals defined up to proportionality from (2.13) though the alternative forms of MCMC can equally well be used.

In order to apply the Gibbs sampler algorithm, the corresponding full conditionals for the intercepts and the regression coefficients can be written as

$$P(\beta_{0j} \,|\, \beta_{-0j}, \underline{Y}, X) \propto \frac{\exp(\beta_{0j} \sum_{i=1}^{n} c_{ij})}{\prod_{i=1}^{n} (1 + \sum_{j=1}^{k-1} \exp(X_i{}^T \beta_j))} \tag{2.14}$$

$$P(\beta_{lj} \,|\, \beta_{-lj}, \underline{Y}, X) \propto \frac{\exp(\beta_{lj} \sum_{i=1}^{n} c_{ij}\, x_{lj})}{\prod_{i=1}^{n} (1 + \sum_{j=1}^{k-1} \exp(X_i{}^T \beta_j))} \tag{2.15}$$

respectively, where j varies from 1 to $k - 1$ and l varies from 1 to m. It is to be noted that if there are k categories and m predictor variables, we have ($k -$

1) (*m* + 1) full conditionals in all (refer (2.14) and (2.15). These full conditionals can be shown to be log concave so samples can be generated, say, by using adaptive rejection sampling scheme of Gilks and Wild (1992).

3. A FEW IMPORTANT TOOLS FOR STUDYING MODEL COMPATIBILITY AND COMPARISON

This section although appears to be an odd combination at first sight, has been brought here for the completeness of the work. The section mostly discusses a few important tools for studying model compatibility and model comparison in Bayesian paradigm. Such tools are numerous in number but we provide a very brief review discussion focusing on only those tools which are used in the latter sections. The interested readers may refer the cited references for a complete discussion and relevant material.

It is to be noted that whenever we assume a model for a given data, it is important to know if the model under consideration is compatible with data in hand. Model compatibility study in Bayesian paradigm can be done in a variety of ways involving both formal and informal approaches. An informal approach can be based on predictive simulation idea where compatibility can be judged by plotting certain characteristics of both observed and predictive data sets, latter obtained on the basis of assumed modelling formulation in Bayesian paradigm. If the plot shows identical behavior of the two data based characteristics, there is no issue to go against the model (a combination of both likelihood and prior) and, as such, it can be considered compatible with the data in hand. This informal approach has been discussed by a number of authors where the authors have advocated considering certain graphical tools for plotting characteristics such as empirical distribution functions or hazard functions obtained separately for both observed and predictive data sets, Upadhyay and Smith (1994). Formal tools based on numerical summaries such as Bayesian versions of p-values can also be used but we do not go into details of such measures as our primary objective does not focus much on the study of model compatibility. One can refer to Bayarri and Berger (1998) for details on such measures.

The model comparison tools, on the other hand, compares two models where both the models happen to be important candidates for the data in hand. It is to be noted that we are mainly interested in testing of statistical hypothesis or variable selection but we visualize this as a problem of model comparison. To provide our exact objective, let us reconsider the polytomous logistic regression model discussed in Section 2. This model is likely to incorporate too many parameters in the form of regression coefficients especially when the situation requires considering a large number of predictor variables. In such a case, interest may often center to see if some of the regression coefficients can be tested against zero. If the conjecture of testing against zero is found to be correct, the dimensionality of the original problem reduces and thereby providing the scope for simplified inferential developments. The problem can be alternatively defined as that of model comparison where we have a model with large dimensionality on one hand and a model with reduced dimensionality on the other.

A number of tools have been proposed for comparing the models in Bayesian paradigm. The most pertinent being the Bayes factor that requires considering the ratio of weighted likelihood functions for the two models where weights are being offered by the considered prior distributions. The Bayes factor is certainly the most appealing measure although it suffers from a few important caveats especially when the priors are vague and the dimensionality of the model demands for extensive and sophisticated computational strategies, Upadhyay *et al* (2012). An easy to use version proposed by Aitkin (1991) is the posterior Bayes factor (PBF) that has been used extensively in the literature although it suffers from an important drawback (see also Upadhyay and Peshwani (2007)). We shall not go into details of such issues rather propose to use the version simply for its computational ease. The posterior Bayes factor can be defined as

$$B_{12} = \bar{L}_1^{po} \,/\, \bar{L}_2^{po}\,, \tag{3.1}$$

where $\bar{L}_i^{po}$ is posterior mean of likelihood under the model i, i = 1, 2. Obviously, one can go with model 2(1) if B_{12} is less (greater) than unity.

The Bayes information criterion (BIC) and the Deviance information criterion (DIC) are the two other important measures that have been considered extensively in the literature. The BIC for model i can be defined as

$$BIC_i = -2 * \log(\hat{L}_i) + k_i * \log(n), \;\; i = 1,\, 2 \tag{3.2}$$

Where $\widehat{L}_i$ is the likelihood calculated at posterior mode although posterior mean is also recommended in the literature. k_i is the number of parameters in the model *i* and n is number of observations. It is to be noted that the second term in BIC is used to penalize a more complex model. The model corresponding to least value of BIC is finally recommended.

Similarly, DIC for model *i* can be defined as

$$DIC_i = \overline{D_i(\Theta_i)} + P_{D_i} \ , i = 1, 2, \ldots \tag{3.3}$$

where $D_i(\Theta_i)$ is the deviance defined as $D_i(\Theta_i) = -2log(L_i(\underline{Y}|\Theta_i))$, $\overline{D_i(\Theta_i)}$ is the posterior mean of deviance, Θ_i denotes the vector valued parameter associated with the model *i*. The second term P_{D_i} in (18) is the number of effective parameters for the model *i*, defined as $P_{D_i} = \overline{D\iota(\Theta\iota)} - D(\hat{\theta}_i)$, where $\hat{\theta}_i$ is an estimate of Θ_i, usually taken as posterior mean, Spiegelhalter *et al* (2002). The term P_{D_i} can also be interpreted as expected reduction in uncertainty due to estimation and, therefore, it is natural to consider it as a measure of model complexi

4. A REAL DATA DESCRIPTION AND THE CORRESPONDING BAYESIAN MODELLING SUMMARIZATION

The section provides a description of data on concentration of bile acid constituents (in mg/ml) collected at SS Hospital, Banaras Hindu University, among n = 61 gallbladder patients. The bile acid constituents are cholic acid (CA), chenodeoxycholic acid (CDCA), deoxycholic acid (DCA) and lithocholic acid (LCA) where the first two acids are referred to as the primary acids and the remaining two as the secondary acids. Bile samples were collected from three groups of patients, namely, control, cholelithiasis and gallbladder carcinoma, admitted to the University hospital. Patients who underwent laporotomy for diseases other than hepatobiliary tract served as control. Patients of benign gallbladder disease who were diagnosed to have gallbladder stone and underwent cholecystectomy and histologically proved to be benign lesion served as the cholelithiasis group and the third group consisted of patients of carcinoma of the gallbladder who underwent laparotomy and subsequently confirmed by histopathology to be malignant.

Each group consisted of 20 patients except the cholelithiasis group that had 21 patients. The complete set of observations are not reported due to confidentiality reasons, however, interested readers may contact the authors for any further query.

The observations were taken by fine needle aspiration of the gallbladder during laparotomy with extreme care to avoid contamination of bile samples with blood. The samples were stored at $-20^{\circ}C$ until analyzed. Peaks of individual bile acids were identified by comparison with peaks of standard bile acids. CA was determined independently and concentration of other constituents was determined according to their ratio to CA from the gas-liquid chromatogram. Among early work on the data, one can refer to Shukla (1993), Makkar (2009). The authors observed that the patients with gallbladder carcinoma had higher concentration of secondary bile acids in comparison to the patients in other two groups. Makkar (2009), however, attempted to draw the same conclusion by considering a particular case of generalized linear model but she worked with only a part of the data and, as such, the sample size was not enough.

Table 2: Classification of four bile acid constituents in accordance with polytomous scheme

Patient no.	CA	CDCA	DCA	LCA	Response variab Y
1	x_{11}	x_{21}	x_{31}	x_{41}	0
2	x_{12}	x_{22}	x_{32}	x_{42}	0
...	...	...	...	...	...
i	x_{1i}	x_{2i}	x_{3i}	x_{4i}	1
...	...	...	...	...	...
n	x_{1n}	x_{2n}	x_{3n}	x_{4n}	2

We, however, visualize the data in a slightly different manner in order that it may appear appropriate for the modelling formulation given in Section 3. We associate with each patient a polytomous response variable Y taking values 0, 1, and 2 where the value 0 corresponds to control group, 1 corresponds to cholelithiasis and 2 corresponds to carcinoma group. As such the data may finally appear in accordance with the pattern shown in Table 2.

Obviously, the scheme reflects the simplified version of polytomous data shown in Table 1 with concentration of four bile acid constituents in the gallbladder as the predictor variable and different gallbladder diseases to a patient as the response variable. It is to be noted that Y takes value 0 for 20 control patients, 1 for 21 cholelithiasis patients and 2 for remaining 20 carcinoma patients.

Obviously, the response variable Y_i with the three possible responses (0, 1, 2) can be converted into a random vector $\boldsymbol{C}_i$ for the patient i, $i = 1, \ldots , n$. The realizations for the corresponding random vector can be given by c_{i1} and c_{i2} and according to the structure given in Table 1, the response variable and the associated random vector for the three categories can be given as follows

Table 3: Polytomous structure for gallbladder disease corresponding to patient *i*

Response variable Y_i	RandomVector $\boldsymbol{C}_i$	
	c_{i1}	c_{i2}
0	0	0
1	1	0
2	0	1

For numerical illustration, we shall consider the same real data set with an objective to examine how the concentration of bile acid constituents effect the three categories of gallbladder diseases.

Before we come to the next section, let us re-write the Bayesian modelling formulation given in Section 2 for $k = 3$ in order to cover the present data description. If (c_{i1}, c_{i2}) denotes the realization of random vector for i^{th} response, the corresponding likelihood function can be written as

$$L(\underline{Y}/\beta_1,\beta_2,X)=\prod_{i=1}^{n}\frac{(\exp(X_i^T\beta_1))^{c_{i1}}(\exp(X_i^T\beta_2))^{c_{i2}}}{1+\exp(X_i^T\beta_1)+\exp(X_i^T\beta_2)} \tag{4.1}$$

Where

$$X_i^T\beta_1=\beta_{01}+\beta_{11}x_{1i}+\beta_{21}x_{2i}+\beta_{31}x_{3i}+\beta_{41}x_{4i}$$
$$X_i^T\beta_2=\beta_{02}+\beta_{12}x_{1i}+\beta_{22}x_{2i}+\beta_{32}x_{3i}+\beta_{42}x_{4i}$$

In (4.1), both β_{01} and β_{02} denote the intercepts associated with control versus cholelithiasis groups and control versus carcinoma groups, respectively. Similarly, β_{11} (β_{12}), β_{21} (β_{22}), β_{31} (β_{32}), β_{41} (β_{42}) are the regression co-efficients associated with bile acid constituents CA, CDCA, DCA and LCA, respectively, corresponding to control versus cholelithiasis (control versus carcinoma) groups and, being the results of logit link, these may be interpreted as the change in log odds for the unit change in corresponding predictor variable when the others remain fixed.

The prior distributions for intercepts and regression co-efficients can be defined independently for each of these parameters and, in lack of any authentic information, the same can be defined as in (2.12) with j = 1 ,2 and l = 0, 1, ... ,4. Moreover, we propose to consider large values for the prior hyperparameters $U_{l\,j}$ and $V_{l\,j}$ in order that the prior remains vague in each case. Thus combining the prior distributions so defined with likelihood in (4.1) via Bayes theorem results in the joint posterior distribution that can be written up to proportionality as

$$P(\beta_1, \beta_2 \mid \underline{Y}, X) \propto \prod_{i=1}^{n} \frac{(\exp(X_i^T \beta_1))^{c_{i1}} (\exp(X_i^T \beta_2))^{c_{i2}}}{1 + \exp(X_i^T \beta_1) + \exp(X_i^T \beta_2)} \qquad (4.2)$$

where the ranges of β_1, β_2 are same as those given in the corresponding prior distributions. Like (2.13), the posterior distribution (4.2) is also difficult to solve analytically and, therefore, one can use Markov chain Monte Carlo simulation to get the desired sample based posterior characteristics. It can be shown that the form (4.2) is a good candidate for Gibbs sampler algorithm with all its full conditionals available from the viewpoint of sample generation. In fact, all the ten full conditionals corresponding to β_{01}, β_{02} and (β_{11}, β_{12}), ... , (β_{41},β_{42}) can be shown to be logconcave and, therefore, adaptive rejection sampling algorithm can be used for sample generation in each case.[11]

4.1 Numerical results on intercepts and regression coefficients

To implement the Gibbs sampler algorithm on the posterior (4.2), we consider a single long run of the chain using least squares estimates of the intercepts and regression coefficients as the initial values.

The convergence of iterating Markovian chain is assessed by monitoring the ergodic averages for each unknown variate in (4.2). The convergence monitoring is successfully achieved at about 50K iterations for each of the unknown variate value although the chain was allowed to run beyond 100K iterations to have an added guarantee on the convergence. Once the convergence is assured, we select samples of size 1K from each of the marginal posteriors of the intercepts and regression coefficients.[12] To minimize serial correlation among the generating variates, selection of final 1K sample is done by picking up variate values at every 10th iteration.

The sample based estimates of a few important posterior characteristics are shown in Table 4. These characteristics are in the form of sample based estimates of posterior mean, median, mode, lower quartile (Q_1), upper quartile (Q_3) and posterior variance obtained for various marginal posteriors. Table also provides the estimated highest posterior density interval (HPDI) with probability coverage 0.95 for each of the model parameters.

The values corresponding to β_{01} and β_{02} represent the estimated posterior intercepts. These values may be of relevance when all the biliary acid constituents are zero, a situation that is almost impossible in any medical finding. Thus the intercepts have no intrinsic meaning and, therefore, we skip any further discussion on their estimated results and it has not been included in the table. It can be further seen that the estimated regression coefficients (β_{11} , β_{21}) and (β_{12}, β_{22}) corresponding to primary acids CA and CDCA are all negative for both cholelithiasis and carcinoma groups when compared with the control group. These results have an obvious reverse interpretation. That is, if the concentration level of primary acid depreciates, a patient in the control group has higher possibility of entering into the diseased group. This interpretation is obviously given under the assumption of changing the concentration of one primary acid and keeping other at the same level. Also, a larger numerical value of regression coefficient of CA than that of CDCA indicates a higher effect of CA as the risk factor than that of CDCA.

The data set also reveals significantly higher concentration of two secondary acids, DCA and LCA, in cholelithiasis and carcinoma groups in comparison to the control group. It can be seen that the positive values of regression coefficients corresponding to secondary acids (β_{31}, β_{41}) and (β_{32}, β_{42}) indicate that the chances of entering into cases from control increase as the levels of secondary bile acids increase. It can be further seen that a larger numerical value of regression coefficient corresponding to DCA than that corresponding to LCA indicates a higher effect of former as the risk factor for carcinogen than that of latter although the trend is reversed for cholelithiasis group (see Table 4).

Table 4: Sample based marginal posterior characteristics of various regression coefficients

Variate	Q_1	**mean**	**median**	**mode**	**Q_3**	**posterior var**	**HPDI**
β_{11}	-0.587	-0.496	-0.481	-0.444	-0.382	0.026	(-0.797, -0.198)
β_{21}	-0.235	-0.156	-0.148	-0.144	-0.063	0.018	(-0.424, 0.097)
β_{31}	0.039	0.172	0.165	0.144	0.292	0.035	(-0.181, 0.548)
β_{41}	0.858	1.277	1.233	1.206	1.625	0.337	(0.219, 2.391)
β_{12}	-9.429	-8.349	-8.776	-9.182	-7.529	2.103	(-9.999, -5.517)
β_{22}	-1.727	-1.359	-1.383	-1.461	-1.008	0.306	(-2.387, -0.252)
β_{32}	6.756	7.653	7.795	7.846	8.723	2.634	(4.146, 10.541)
β_{42}	0.456	1.961	1.889	1.732	3.496	5.098	(-2.035, 6.816)

The marginal posterior density estimates of regression coefficients β_{11}, β_{21}, β_{31} and β_{41} are shown in Figure 1 in the form of histograms whereas the corresponding density estimates of β_{12}, β_{22}, β_{32} and β_{42} are presented in Figure 2. The vertical line in each Figure corresponds to maximum likelihood estimate of the corresponding model parameter. Obviously, the figures provide an overall impression about different marginal densities of various regression coefficients. Say, for instance, the parameters β_{11} and β_{21} are both negatively skewed with small variability in the two cases and the values mostly inclined towards negatives, although the value of β_{21} is very much close to zero, a conclusion that is obvious from Table 4 as well. The parameters β_{31} and β_{41}, on the other hand, are almost symmetrical, the former having high probability for a value close to zero whereas the latter having high probability for a value close to unity (see also Table 4). Moreover, the estimated posterior variability is also small in case of β_{31} with an overall impression that the values are clustered around zero (see Figure 1). Similarly, the parameters β_{12} and β_{32} are skewed with former having positive skewness and the latter having negative skewness. The parameters β_{22} and β_{42}, on the other hand, appear to be more or less symmetrical. It is to be noted that maximum likelihood estimates can also be considered equally well in the situations where estimated posterior densities appear to be more or less symmetrical. On the other hand, skewed situations usually recommend posterior modes as the appropriate point estimates.

The density estimates shown in Figures 1-2 along with the estimated interval estimates for various regression coefficients provide an impression that at least the parameters β_{21} and β_{31} may be tested against zero. If such a testing results in a conclusion that supports the conjecture to take a value zero for a particular parameter, the resulting model is going to be simplified. We may then proclaim that we are not losing anything by using a reduced model (say, M_2 or M_3) over the full model (say, M_1) where the models M_2 and M_3 are defined after removing the parameter β_{21} and β_{31}, respectively. In other words, we can say that the corresponding biliary acid constituent can be considered to have almost no affect in the development of the disease (cholelithiasis). This issue may equally well be treated as the problem of model comparison where a simplified model may be compared with the full model and accordingly the conclusion may be looked upon.

The complete analysis of the model M_2, M_3 and the interpretation of the estimates of corresponding model parameters can be done similarly as it was discussed with the estimated values in Table 4. We,

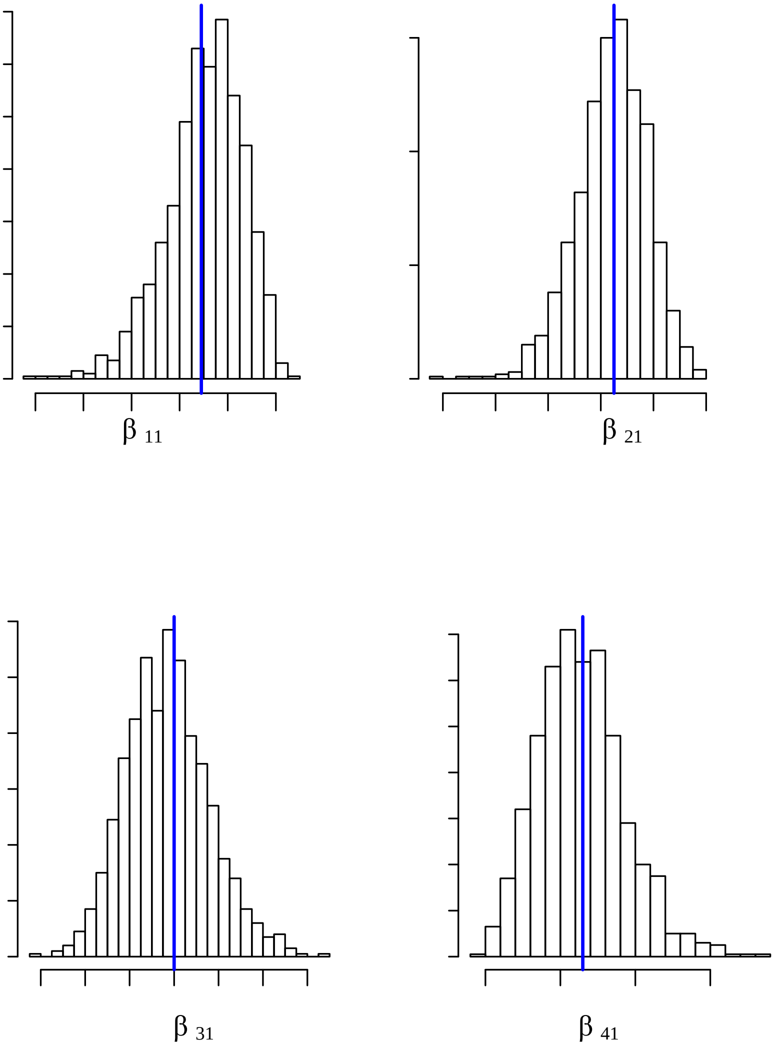


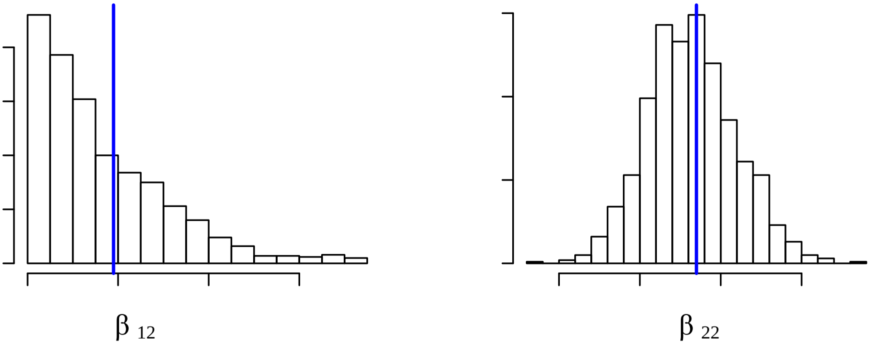
Figure 1: Histogram showing the posterior density estimates of β_{11}, β_{21}, β_{31} and β_{41}.

however, skip this discussion noticing that there is nothing specific to offer beyond what has been discussed based on Table 4. The results for the parameters of these models (M_2 and M_3) were found to be more or less similar to those obtained with the corresponding parameters in M_1.

5. MODEL COMPATIBILITY AND COMPARISON

The aim of this section is to study compatibility of the full model M_1 with the considered data set and then to provide a comparison of the full model M_1 with the reduced models M_2 and M_3. We shall use the informal strategy for model compatibility based on predictive simulation ideas briefly discussed in Section 3. Our informal approach involves comparing the empirical distribution function (Edf) plots corresponding to the observed and the predicted data sets obtained from the model under consideration (see (4.1)-(4.2).

For our intended purpose, we initially considered the Edf plot corresponding to the observed data, shown in Figure 3 in the form of solid line. We next considered 25 predictive samples, each of size similar to that of observed data. The Edf plots corresponding to predictive data sets are superimposed as dotted lines in Figure 3. Although it is an informal approach, it clearly provides a message that there is no discrepancy



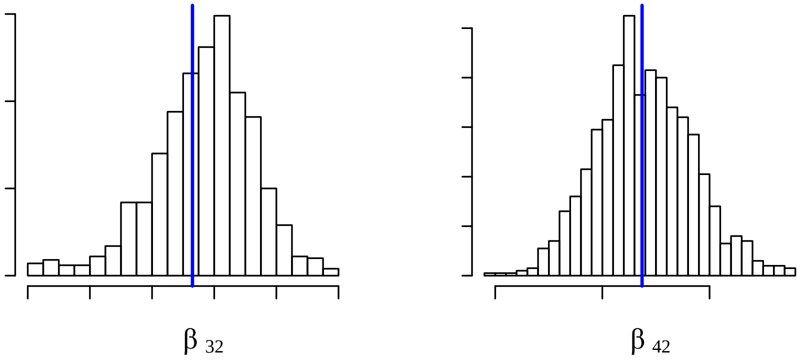


Figure 2: Histogram showing the posterior density estimates of β_{12}, β_{22}, β_{32} and β_{42}.

among the observed data based Edf plot and the model based predictive data plots. Obviously, the model M_1 can be convincingly used for the data in hand.

Results of the comparison of models M_1, M_2 and M_3 are shown in Table 5 in the form of BIC, DIC and PBF values (see Section 3). It is obvious that all the three values support the reduced models although the extent of support is not appreciably enough. Since the parsimony principle also conveys going with the simpler model, it is suggested to consider the reduced model in spite of the fact that the BIC and the DIC values are quite close to each other for the considered models and the PBF value is also seen to be not far away from unity. We can, therefore, safely say that the reduced models M_2 and M_3 can be recommended for the data in hand and simultaneously the biliary acid constituent CDCA and DCA separately do not appear to have an appreciable role in the development of cholelithiasis. It is to be noted that result of simultaneous testing of β_{21} and β_{31} against zero is not given in Table 5 as it gave worse model then the full model.

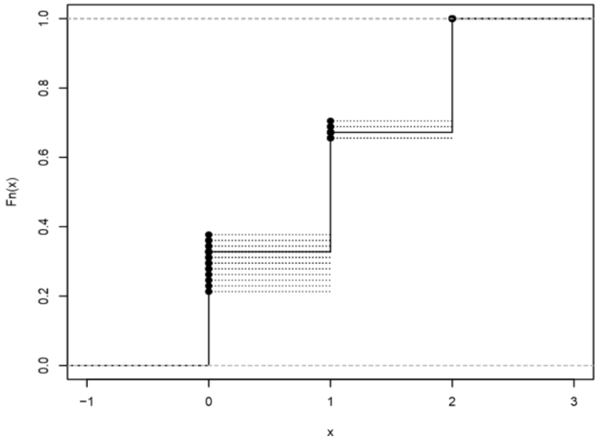


Figure 3: Edf plots corresponding to the observed and predictive data sets.

Table 5: Values of BIC, DIC and PBF corresponding to the models M_1, M_2 and M_3

Model	BIC	DIC	PBF
M_1	70.81	41.97	–
M_2	67.26	39.86	0.81
M_3	68.84	39.92	0.85

6. CONCLUSION

This is an extensive study on multivariate extension of generalized linear model to cover the case of polytomous data. The study is exclusively Bayesian based on proper vague priors for the parameters involved in the model. The novel feature of the study includes the analysis of a real data set on the gall bladder patients. As pointed, this study has been considered earlier by a number of authors but mostly based on some unrealistic assumptions or some simplified tools to avoid statistical complications. Our finding reveals that the levels of CA are significantly lower in the cholelithiasis and gallbladder carcinoma group as compared to the control group. The two secondary acids, namely DCA and LCA, which are normally present in small quantities in bile, are found to be significantly higher in concentration in carcinoma patients. Bile constituents CDCA and DCA do not have any significant role in developing gallstone. Our analysis, however, reveals that the roles of CA and DCA may be considered as possible carcinogen. Literature also suggests that the bacterial degradation of primary acid is ultimately responsible for the formation of secondary acid which are tumor promoters.

REFERENCES

Agresti A. (2002): *Categorical Data Analysis, 2nd ed.*, New York: Wiley.

Aitkin M. (1991): Posterior Bayes Factor. *Journal of the Royal Statistical Society. Ser. B*; **53**(1): 111-142

Andrich D. (1978): A rating formulation for ordered response categories. *Psychometrika* 1978; **43**: 561-573.

Bayarri MJ and Berger JO. (1998) Quantifying surprise in the data and model verification. In *Bayesian Statistics, eds.; 53-83*. J.M. Bernado, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Oxford University Press, London.

Congdon P.(2004): *Applied Bayesian Modeling*, New York: Wiley

Draper NR and Smith H. *Applied Regression Analysis, 3rd ed.,* 1998. New York: Wiley.

Gelman A and Hill J. (2007): *Data Analysis Using Regression and Multilevel/Hierarchical Models,* Cambridge University Press.

Gilks WR and Wild DG. (1992): Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**: 337-348.

Liang KY. (1999): *Generalized linear models, estimating functions ad multivariate extensions,* Department of Biostatistics, School of Hygiene and Public Health, Johns Hopkins University.

Makkar P. (2009): *Bayesian Solutions of Some Medical Data Problems*, Unpublished Ph.D. thesis, Department of Statistics, Banaras Hindu University, India.

McCullagh P and Nelder JA. (1989): *Generalized Linear Model, 2nd ed.*, 1989. Chapman & Hall.

Nelder JA and Wedderburn RWM. (1972): Generalized liner models. *Journal of the Royal Statistical Society*; **135**(3): 370-384.

Rasch G. (1961) On the general laws and the meaning of measurement in psychology. In J. Neyman (1961): (Ed.), *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, **4**: 321333. Berkeley: University of California Press.

Shukla VK, Tiwari SC and Roy SK. (1993): Biliary bile acids in cholelithiasis and carcinoma of gall bladder. *European Journal of Cancer Prevention*; **2**:155-160.

Spiegelhalter DJ, Best NG, Carlin BP and Vander Linde A. (2002): Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*; Ser. B, **64**: 583-640.

Tuerlinckx F and Wang WC. (2004): Models for polytomous data. In *Explanatory item response models*; 75-109. Springer, New York.

Upadhyay SK and Smith AFM. (1994): Modelling complexities in reliability, and role of Bayes simulation. *International Jr. Cont. Eng. Educ.: Sp. Issue on Applied Probability Mod.*; **4**: 93-104.

Upadhyay SK and Peshwani M. A (2007): Bayes analysis of Birnbaum-Saunders distribution using the Gibbs sampler approach. Upadhyay SK, Singh U and Dey DK, (2007): eds., *Bayesian Statistics and Its Applications*: 438-455. Anamaya, New Delhi,

Upadhyay SK, Gupta A and Dey DK.(2012): Bayesian modeling of bathtub shaped hazard rate using various Weibull extensions and related issues of model selection. *Sankhya: The Indian Journal of Statistics*; **74-B**(1): 15-43.